Název práce: Analýzy reálných dat a jejich využití

Autor: Jakub Stárka

Katedra: Katedra softwarového inženýrství

Vedoucí disertační práce: RNDr. Irena Holubová, Ph.D.

Abstrakt: Znalost reálných dat je základem pro optimalizaci mnoha technik zpracování dat. Jejich získání, analýza či integrace zahrnují mnoho problémů, na které je zaměřena tato práce. Mezi tyto hlavní problémy patří např. automatické stahování dokumentů, extrakce dat a jejich analýza, či odvozování schémat.

V této práci popíšeme komplexní framework, který umožňuje opakovaně provádět statistickou analýzu nad reálnými XML dokumenty, které jsou získané z internetu. Také navrhneme několik charakteristik pro XML dokumenty, RDF trojice a XQuery dotazy včetně podrobných výstupů analýz nad několika veřejně dostupnými kolekcemi dat. V neposlední řadě popíšeme rozšiřitelný nástroj pro odvozování XML schémat. Díky jeho modulárnímu designu je možné kombinovat několik nezávislých přístupů pro jednotlivé kroky. V rámci práce nepopíšeme jen samotný framework, ale i oblast odvozování jako takovou a s ní související problémy.

Klíčová slova: analýza dat, extrakce dat, odvozování schémat